# The length factor in automatic bilingual terminology extraction

Volk, Martin ; Pantli, A-K ; Malka, A M

Abstract: This paper describes an evaluation of filtering methods for bilingual terminology extraction. Terminology extraction systems often favor recall over precision. This strategy results in an enormous number of term candidate pairs that have to be manually checked and cleaned. In the most extreme the post-editing step is so cumbersome that it prevents a system from practical employment. We show that filters based on formal criteria efficiently help in reducing manual labor. The most promising filter is based on the length difference between the source term candidate and the target term candidate.

In *Proc. of 6th International Conference on Terminology and Knowledge Engineering.* Nancy. 2002

# The length factor in automatic bilingual terminology extraction

**Martin Volk, Anna-Katharina Pantli, Anita Mirjam Malka**

Zurich University of Applied Sciences Winterthur*
Department of Applied Linguistics and Cultural Studies
Thurgauerstrasse 56, CH-8050 Zurich
{vlk,pnt}@zhwin.ch

**Abstract**

This paper describes an evaluation of filtering methods for bilingual terminology extraction. Terminology extraction systems often favor recall over precision. This strategy results in an enormous number of term candidate pairs that have to be manually checked and cleaned. In the most extreme the post-editing step is so cumbersome that it prevents a system from practical employment. We show that filters based on formal criteria efficiently help in reducing manual labor. The most promising filter is based on the length difference between the source term candidate and the target term candidate.

## 1. Introduction

In recent years, terminology management has become increasingly important for many international companies. They have come to realize that terminology management means knowledge management and is therefore of critical importance for their business. Terminology has to be collected, administered, maintained and disseminated. Terminology database systems have been on the market for some time and they support the latter phases. But the collection of terms, in particular manual term extraction, is very time-consuming and error-prone. For this reason, there is considerable interest in automatic term extraction. For some years, terminology extraction (TE) tools have been developed in research laboratories (for an overview see Bourigault et al. 2001); some of them have recently entered the market. These tools analyse texts linguistically and/or statistically and work with either monolingual or bilingual corpuses. However, automatic term extraction is difficult and current tools are far from perfect.

In the technical literature the topic has been discussed from various points of view and under differing headers. A classic is the article by Dagan et al. (1993) which introduced statistical means for bilingual word alignment. Later work often focused on the detection of terms and how to discriminate them from non-terms. Approaches vary from linguistics (e.g. Heid et al. 1996; Heid 1999; Jacquemin 2001) to purely statistical methods based on measures like feature frequency and inverse document frequency that are well known from information retrieval (see Kageura and Umino 1996).

## 2. Goal of the paper

In this paper we explore a number of filter heuristics for reducing manual post-editing of automatically computed term candidate pairs. The idea is to exploit the formal properties of source term candidates and target term candidates in order to automatically filter and thus reduce the number of incorrect term candidate pairs. We will show that a filtering heuristic based on the length difference is the most promising.

## 3. The Term Extraction tool

In this project we used a commercial TE tool that was made available to us for evaluation purposes. The TE tool works in the following manner: When processing monolingual texts, the tool linguistically analyses each sentence of the source text. The words are morphologically processed and the resulting set of readings is disambiguated with respect to word classes with a part-of-speech tagger. Then, all noun phrases (NPs) in the source text are extracted by using patterns over parts-of-speech. Thus, the user receives a list of term candidates, from which he or she then chooses the terms.

When working with bilingual (parallel) texts, the system aligns the two texts before running the extraction. It links each sentence of the source text with a sentence of the target text. This procedure is approximative because, on the one hand, the system does not always properly recognise the sentence final punctuation and, on the other hand, texts are not always translated sentence by sentence. The user is asked to correct alignment mistakes manually before running the term extraction.

Bilingual term extraction is performed in two steps. First, all noun phrases of a corresponding sentence pair

* This research was partly funded by CTI (Swiss Commission for Technology and Innovation), CTI Project Nr 5292.1 FHS

are identified as term candidates. Second, all term candidates in the source unit are linked to all term candidates in the target unit. This leads to an enormous number of term candidate pairs and to countless pairs that have to be eliminated or corrected manually. Let us assume there are 4 terms in the source unit and 4 terms in the target unit. Then the system will generate 16 term candidate pairs, out of which only 4 are correct correspondences. This proliferation of term candidates leads to a very labour-intensive post-editing. Therefore, we have implemented a number of filter heuristics to increase the term extraction precision without (much) loss in recall.

We selected two technical manuals (each with English as source language and German translation) and semi-automatically determined the terms contained in these documents. We used these terms as a gold standard for comparison. We will call the two manuals D1 and D2 in this report.

## 4. The Procedure for defining the gold standard

Our extraction procedure started with monolingual term extraction from both source texts. The English source texts D1 (4874 words)[1] and D2 (11,556 words) were processed by the term extraction (TE) system in order to extract the term candidates and list them in two databases. Term candidates are single nouns, noun sequences, and adjective-noun NPs (cf. to categories found by Justeson, and Katz (1995)). D1 resulted in 756 term candidates and D2 in 1351 term candidates. Both lists were then checked manually, and all term candidates were labelled with a grade between 1 (not a term) and 5 (definitely a term). The grades were defined as follows:

1. The candidate is not a term in the given subject domain (it is an item of general vocabulary).
2. The candidate is probably not a term. It is used in the text with a special meaning but it is otherwise unusual in the subject domain.
3. The candidate is/was a term in the subject domain, but it is also an item of general vocabulary with a similar meaning, or it is unclear whether the candidate has a special meaning in the given text.
4. The candidate is probably a term. It could be a compound with composite semantics that does not require a separate entry of the term in the database.
5. The candidate is definitely a term. It does not occur in general vocabulary or it only occurs in general vocabulary with a different meaning.

These graded lists were then counterchecked against the source texts in order to find terms missed by the program. When we added all term candidates graded 2 through 5 and all manually added terms, we had lists of 133 "good" terms for D1 and 350 "good" terms for D2.

The second step was to have the TE system process both source and target texts D1 and D2 for bilingual term extraction. After automatic alignment, the sentence pairs were manually checked and, where necessary, corrected. Then, the system generated bilingual lists of term candidates, which resulted in 3950 pairs for D1 and 6498 pairs for D2.

These bilingual pairs were automatically checked (with a Perl script) against the manually checked monolingual list of English terms. Only candidate pairs that had an English entry in the monolingual list were kept. In other words, all candidate pairs with invalid English terms were eliminated, and the number of terms thus reduced to 919 term pairs for D1 and 1933 term pairs for D2.

These lists were again corrected and cleaned manually. This led to another reduction in the number of term pairs (109 for D1 and 258 for D2). We used these lists as the gold standard.

The task then was to develop heuristics that would get us from the thousands of automatically computed bilingual term candidate pairs as close as possible to the few hundred true term pairs.

## 5. Problems in term judgements

Before we come to the heuristics we want to briefly comment on the problems encountered in term judgements during manual correction. Correcting the lists was not always easy. Since the TE system treats different languages in slightly different ways, term pairs very often matched only partially. The bilingual lists also showed that sometimes the translation lacks consistency:

*Copy contrast* → *Kopien-Schwärzungsgrad*
*Copy contrast* → *Schwärzungsgrad*
*Copy contrast range* → *Schwärzungsgrad*

Moreover, the German target text did not always provide an exact equivalent to the English term, making the TE system incapable of matching these terms correctly. In such cases, of course, the program cannot generate term candidates that are better than the texts provided.

Our strategy for the problematic cases was as follows:

- If the term candidate pair did not contain a satisfactory term in the target language, the pair was eliminated. This accounts for the difference between the number of monolingual terms and bilingual term pairs. If source expression and target expression matched only partially, superfluous material was eliminated (numbers, attributes) or missing material inserted.
- If the source expression constituted a good term but the target expression a bad term, then we tried to optimise it. We never substituted the target expression by a completely new term. The gold standard is meant to reflect the translations in the underlying texts and not our terminological preconceptions.
- Inconsistencies were preserved. If, for instance, the English term X was translated by both German Y and Z in the underlying texts, both pairs X → Y and X → Z were accepted.
- A few terms that were not extracted as such by the TE system were added (e.g. *tab* → *Registerkarte*).

[1] All word counting was done in MICROSOFT WORD 2000.

In retrospect we think that we were too stringent in our judgements, so that the resulting list represents the true core of terms but for practical purposes more candidate pairs might be acceptable terms.

## 6. Precision and Recall of Automatic Term Extraction

We can now compare the automatic lists of term candidate pairs as computed by the TE system to our gold standard lists. We accept a candidate pair as correct if either the base form pair or the surface form pair is in the gold standard list. For text D1, we manually determined 109 correct term pairs. 76 of these are in the automatic list of term candidate pairs. The number of pairs in that list is 3950. So we observe a precision of 1.9% and a recall of 69.7%. For text D2, the gold standard list consists of 258 term pairs. 208 of these pairs are in the list of automatic term candidate pairs. That lists comprises 6498 entries. This corresponds to a precision of 3.2% and a recall of 80.6%.

Most of the missing terms from the gold standard list (sometimes called "silence") can be traced back to some special cases:

1. If a verb is part of either the German or English term, then the noun phrase (NP) is often not correctly extracted.
   *Paper Tray Selector → Taste Papierkassette wählen*
2. Sequences of German nouns are not recognized as one NP. This is a serious shortcoming since such sequences are frequent in technical texts.
   *configuration tab → Registerkarte Konfiguration*
3. Prepositional phrases (PPs) are not recognized as noun attributes. PP attachment is a difficult problem in natural language processing. Fortunately, it occurs in only few cases in terminology extraction.
   *landscape page → Seite mit Querformat*
   On the positive side, we noticed that the TE system recognizes German PPs in prenominal adjective phrases, which is the desired behaviour in terminology extraction.
   *mit Niederspannung betriebene Geräte [low voltage equipment]*

## 7. Heuristics

Heuristics are rules of thumb that lead to the desired result in most cases. The employment of heuristics takes into account that they may sometimes fail. We suggest heuristics to filter term candidates. They can be used to automatically eliminate term candidates or to mark them for manual inspection.

In order to filter bilingual term candidates, we can think of two principally distinct sets of heuristics.

1. Heuristics that are based on the properties of one term candidate in either source or target language.
2. Heuristics that are based on a comparison of the properties of the two words in a bilingual candidate pair.

In addition we have the choice of basing the heuristics on the surface forms of the term candidates (the word forms as they occur in the text) or on the base forms provided by the TE system's lemmatizer.

In this project, four different filtering heuristics were implemented (in Perl) and evaluated. They were based on the observations made during work with the TE system. The heuristics state that term candidates should be eliminated if

1. the source expression contains a number and the target expression does not (or vice versa).
2. the first character of a term candidate is a special symbol.
3. an expression with only one element is paired with an acronym composed of several letters.
4. a short expression is paired with a long one with a length difference exceeding a certain threshold.

### 7.1. Application of the Heuristics

Each line in the automatically generated bilingual term candidate file contained the 5-tuple as can be seen in table 1.

| English base form | English surface form | Co-occurrence frequency | German base form | German surface form |
|---|---|---|---|---|
| *cartridge* | *cartridge* | 1 | *Patrone* | *Patrone* |
| *cartridge* | *cartridge* | 1 | *alt Tonerpatrone* | *alten Tonerpatrone* |
| *cartridge* | *cartridge* | 2 | *grün Hebel* | *grünen Hebel* |
| *cartridge* | *cartridge* | 2 | *Tonerfüllung* | *Tonerfüllung* |
| *cartridge* | *cartridge* | 2 | *Kopie* | *Kopien* |

Table 1: Bilingual term candidate files, example.

We used a two-step procedure to make the application and the evaluation of the heuristics as transparent as possible. The first step was to use the heuristics to annotate all term candidate lines with a judgement. The algorithm works as follows:

The program steps through all lines of automatically extracted term candidates. It annotates every line with a judgement about the quality of the term in the following order:

1. If the English base form or the German base form starts with any of the symbols '*.%/', then the line is annotated as 'Garbage start symbol'.
   ** use label → Laserdrucker [laser printer]*
2. If the English surface form contains a number and the German surface form does not (or vice versa), the line is annotated as 'Number difference'.
   *inch floppy disk drive → 1.44 MB*
3. If the English base form is an acronym (a sequence of two or more capital letters) and the German base form is a non-compounded word[2] (without hyphen and without an intervening blank), then the line is

[2] The lemmatizer built into the TE system supports decompounding and marks compound boundaries.

annotated as 'Acronym difference'. This also works vice versa.
*kit → EEA*

4. If the length difference between the English base form and the German base form (defined as the number of characters) is greater than 10, the line is annotated as 'Length difference'.
*laws → mit Niederspannung betriebene Geräte [low voltage equipment]*

For the document D1 we obtained the automatic judgements in table 2:

| Heuristic | Number of lines | Percentage |
|---|---|---|
| Garbage start symbol | 3 | 0.08 |
| Number difference | 282 | 7.14 |
| Acronym difference | 78 | 1.97 |
| Length difference | 1064 | 26.94 |
| No judgement | 2523 | 63.87 |

Table 2: Automatic judgements for D1

And for text D2 we obtained the automatic judgements in table 3.

| Heuristic | Number of lines | Percentage |
|---|---|---|
| Garbage start symbol | 46 | 0.71 |
| Number difference | 466 | 7.17 |
| Acronym difference | 50 | 0.77 |
| Length difference | 1764 | 27.15 |
| No judgement | 4172 | 64.20 |

Table 3: Automatic judgements for D2

The results are thus relatively consistent over the two texts. Using the four heuristics, we can eliminate around a third of the automatic term candidate pairs. The length-difference heuristic is by far the most influential, followed by the number-difference heuristic. The other two are marginal with respect to their impact on the number of term candidates.

We now have to check whether the heuristics eliminated any term candidates that should have been kept.

## 7.2. Evaluation of the Heuristics

We evaluated the results of the heuristics by comparing the term candidate lines with the lines in the gold standard lists. The evaluation algorithm worked for base form pairs and surface form pairs. In the first round we searched for the exact same pairs. Using the filter heuristics, we cut down the number of term candidate pairs from 3950 to 2523 for D1. It turned out that out of the 76 correct terms left for example text D1, 7 were incorrectly eliminated by the length-difference heuristic, leaving 69 correctly found terms. This corresponds to 2.7% precision (with respect to the reduced term candidate set) and 63.3% recall. Precision still seems intolerably low but a 30% reduction in the term candidate set is surely a significant achievement.

The figures are very similar for text D2. The term candidate set was reduced from 6498 to 4172. Out of the 258 gold standard terms, 50 were missed by the automatic extraction. 11 terms out of the 208 left were incorrectly suppressed by the length-difference heuristic, and 1 term was suppressed by the acronym-difference heuristic. This leaves 196 correct terms (4.7% precision and 76.0% recall).

When we relax the comparison criterion somewhat and allow for partial matches, the results are slightly better. That means we now accept a pair (X,Y) as correct if the manually determined term X is part of an automatically computed term candidate auto-X and if Y is part of the corresponding auto-Y; capitalization is also ignored. The idea is that sometimes terms were manually shortened or corrected, and a full form or slightly deviating form is still useful. This type of comparison leads to 74 correct terms for text D1 and 203 correct terms for D2.

As mentioned above, the length-difference heuristic is a powerful instrument to filter out erroneous term candidates. This observation gave rise to the questions: What happens if we use other length-difference values? Could we have avoided the suppression of correct terms if we had set the difference threshold to 12 or 14? How many term candidates would we have filtered with these difference values?

To provide a clearer picture, we have plotted the behaviour of the length difference over all automatically computed term candidates in the graphs (table 4).

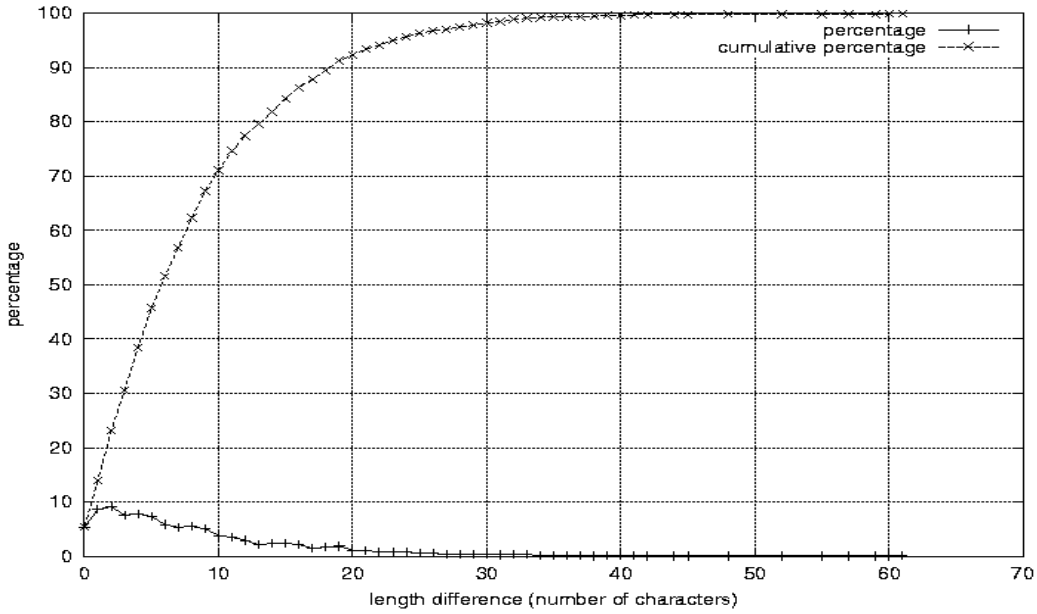


Table 4: Percentage of term candidates, which can be eliminated by a certain length difference

Most interesting is the cumulative percentage curve. It shows what percentage of term candidates can be eliminated by what length difference. If, for instance, we want to eliminate 50% of the term candidates, we need to eliminate all term candidate pairs with a length difference greater than 6 characters. Of course, this will eliminate some wanted pairs. A length difference of 13, on the other hand, would only have eliminated 20% of the pairs. So, the length difference cut-off of 10 used in our heuristics constitutes a good compromise.

Unlike the automatic term candidate lists, our gold standard lists contain only few term pairs with large length differences. These are the ones that are erroneously suppressed by the length-difference heuristic. Some of these pairs are correct translations in which the term of one language is much more concise than the other.

*Settings → Individuelle Einstellungen*

But most are rather strange translations found in the example documents, and we have doubts about accepting them as entries for a terminology database (which is reflected in low grades in our gold standard lists).

*screen → Bildschirmanzeigen [screen display]*

These observations indicate that the length-difference heuristic is very useful, even though it leads to a 4-6% loss in recall in our experiments. It helps to detect mistakes in automatically or manually corrected term lists and should be added to TE systems at least as a sorting criterion. If the term candidate pairs are sorted according to decreasing length difference, the user can easily determine the level to start manual inspection.

## 8. Conclusions

Automatic term extraction is becoming a useful tool for translators and will be an indispensable part of the translation workflow within a few years. But in order to live up to their potential, TE systems need to be improved to find their market position.

Monolingual term extraction as described above followed by manual filtering is surely faster than manual term extraction. And it provides additional information (in particular context information) that is bothersome for manual term extraction. But bilingual term extraction as implemented in our test system results in an overwhelming number of term candidate pairs. The noise in these data makes it difficult if not impossible to use.

We have shown that heuristics can be used to constrain the number of erroneous term pairs. The most powerful is the length-difference heuristic that can easily cut out about a third of the term candidate pairs without much loss in recall.

From our experiences with the TE system we have compiled a number of suggestions to improve such tools.

1. It is of utmost importance that the noun phrase recognition is identical in both languages involved in bilingual term extraction. If the NPs in one language include numbers, then the NPs in the other language must also include numbers. For English it is important that the recognition of nominal compounds is optimised so that these noun sequences match with German compounds that are orthographic units (or hyphenated).

2. Making NP recognition transparent to the user will be very helpful. This should be done by a detailed description of the NP rules (for each language) in a user guide or at least in online help. This will make clear which NPs the system is unable to find, and thus serve to direct the user in manual term checking.
3. All heuristics should be optional, so that the user can choose himself which of them he or she wants to activate.
4. It should be possible to eliminate or mark all words that are items of general vocabulary. In another project we found out that the 10% most frequent words of the source text minus general vocabulary constitute very good term candidates.

Some of the manual annotations that we produced in this project could not be exploited because of time constraints. For instance, the term candidates were labeled with grades on a scale ranging from 1 to 5. This grading, however, was not taken into account in the course of the project. It will be worthwhile to investigate whether this grading corresponds to formal criteria that could be used to filter term candidates. In the same direction, we would like to have a closer look at the partial correspondences between English-German term pairs and compare this to other language pairs. In the automatically extracted lists, many partial matches could be found, such as

*printer driver configuration tab → Konfiguration*
*printer driver configuration tab → Registerkarte*

Such matches were marked at the beginning of the project, but they have proved too numerous, and the idea was dropped.

More heuristics need to be explored. The following heuristic was also considered, but were unable to test it: If a term candidate pair (X,Y) has a high co-occurrence frequency, and all other pairs with X have a significantly lower frequency, they can be eliminated or marked. Of course, this heuristic requires large document collections as a basis for the co-occurrence values and the notion of "significantly lower frequency" needs to be statistically founded or at least empirically proven.

## 9. References


Bourigault, D., C. Jacquemin, M.-C. L'Homme, 2001. *Recent Advances in Computational Terminology*. Natural Language Processing (vol. 2). Amsterdam, John Benjamins.

Dagan, I., K. W. Church and W. A. Gale. *Robust Bilingual Word Alignment for Machine Aided Translation*. Proc. of the Workshop on Very Large Corpora: Academic and Industrial Perspectives, Ohio State University, Columbus, Ohio, July 1993.

Heid, Ulrich, 1999. *Extracting terminologically relevant collocations from German technical texts*. In: Proc. of 5th International Congress on Terminology and Knowledge Engineering, TKE-99: Section on Language Engineering and Terminology.

Heid, Ulrich, S. Jauss, K. Krüger, A. Hohmann, 1996. *Term extraction with standard tools for corpus exploration*. In: Proc. of 4th International Congress on Terminology and Knowledge Engineering, TKE-96. 26-30.

Jacquemin, Christian, 2001. *Spotting and Discovering Terms through Natural Language Processing*. Cambridge, MIT Press.

Justeson, J.S., S.M. Katz, 1995. *Technical terminology: some linguistic properties and an algorithm for identification in text*. In: Natural Language Engineering, 1(1). 9-27.

Kageura, Kyo, Bin Umino, 1996. *Methods of automatic term recognition: A review*. In: Terminology. International Journal of Theoretical and Applied Issues in Specialized Communication. 3(2). 259-289.

L'Homme, M.-C., L. Benali, C. Bertrand, Patricia Lauduique, 1996. *Definition of an evaluation grid for term-extraction software*. In: Terminology. International Journal of Theoretical and Applied Issues in Specialized Communication. 3(2). 291-312.